Review of Master's Thesis

Student: Šulák Ladislav, Bc.
Title: Detection of Malicious Websites using Machine Learning (id 20487)
Reviewer: Černocký Jan, doc. Dr. Ing., UPGM FIT VUT

1. **Assignment complexity** — **more demanding assignment**
   Podle mého názoru náročné zadání, vyžadovalo studium možných útoků pomocí Java script kódu a technik strojového učení (machine learning, ML).
2. **Completeness of assignment requirements** — **assignment fulfilled**
   Zadání bylo splněno a student vykonal značný objem práce na definici parametrů a testování mnhoa metod strojového učení. Poslední bod zadání 6. (analýza) by si ale zasloužil více pozornosti - výsledky jednotlivých ML metod jsou uváděny bez hlubší analýzy proč ta která funguje lépe či hůře, není provedena analýza podle různých typů škodlivého kódu či analýza falešných alarmů u validních stránek atd.
3. **Length of technical report** — **exceeds requirements**
   Rozsah práce je značný a překračuje limity vyžadované fakultou, ale je nevyvážený - zatímco je velká plocha věnována úvodu k různých typům útoků (tato část má pedagogickou hodnotu), směrem ke konci se kapitoly tenčí a na důkladnou analýzu výsledků již zřejmě nezbyl čas a síly.
4. **Presentation level of technical report** — **88 p. (B)**
   Práce je slušně strukturována, některé části se ale opakují a místy by prospěla re-strukturace (např. při uvádění dělení dat na trénovací a evaluační sety).
5. **Formal aspects of technical report** — **85 p. (B)**
   DP je psána slušnou angličtinou, s nenulovým, avšak akceptovatelným množstvím gramatických chyb a překlepů. Tabulky jsou slušně provedeny. Grafy Bohu žel trpí velmi malými popisky os a legendami do té míry, že jsou těžko interpretovatelné, to je u jinak kvalitní práce škoda. Pro čtenáře méně se orientující v bezpečnosti by bylo vhodné připojit slovníček zkratek.
6. **Literature usage** — **90 p. (A)**
   Příkladná ve studiu pramenů ohledně počítačové bezpečnosti, o něco slabší u pramenů týkajících se strojového učení, nejsem si jistý, zda student plně chápal detaily ML algoritmů, které testoval.
7. **Implementation results** — **90 p. (A)**
   Sada SW, především v pythonu, pro crawling a zpracování dat, vlastní ML a testování. Velmi kladně hodnotím moduly pro data engineering - zde je zřejmá erudice studenta i výborná znalost požadavků uživatelů a metod útoků. U vlastního ML se jednalo o použití standardních knihoven, které bylo méně náročné.
8. **Utilizability of results**
   Práce je využitelná pro detekci nového malwaru, pro průmyslové nasazení bude zřejmě nutné další testování a optimalizace, ale je perspektivní.
9. **Questions for defence**
   1. Jaké jste u webových stránek, resp. Java script kódu z nich, využíval labelování - pouze malicious/benign, nebo více kategorií ?
   2. Popište přesněji (např. schématem) vektorizaci na vstupu LSTM.
   3. Komentujte, zda bylo pro všechny dokumenty možné extrahovat všechny parametry, nebo jste některé musel pokládat za "missing features".
10. **Total assessment** — **86 p. very good (B)**
    Slušná práce, které by prospělo větší zamyšlení nad experimentálními výsledky a jejich důkladná analýza. Technická zpráva je čitelná, ale byl by vhodný větší důraz na použité techniky ML a lepší rozmyšlení logického členění práce.

In Brno 7. June 2018

.....
signature